1999

Benchmarking the Advanced Search Interfaces of Eight Major WWW Search Engines

John W. Felts
Coastal Carolina University, jfelts@coastal.edu

Randy D. Ralph
rdralph@gmail.com

Benchmarking the Advanced Search Interfaces of Eight Major WWW Search Engines

Dr. Randy D. Ralph & John W. Felts, Jr.

Keywords: information retrieval, search engines, World Wide Web, benchmarking, advanced search, search interfaces

Abstract: This research project was designed to benchmark the performance of the advanced search interfaces of eight of the major World Wide Web (WWW) search engines, excluding the meta engines. A review of the literature did not find any previous benchmarking studies of the advanced interfaces based on quantitative data. The research was performed by fifty-two graduate students of library and information studies (LIS) on three campuses of the University of North Carolina (UNC) as a class research project for course LIS 645, Computer-Related Technologies in Library Management. The class was offered by the Department of Library and Information Studies at UNC Greensboro through the North Carolina Research and Education Network (NC-REN). The LIS students selected Altavista, Excite, Go/Infoseek, Google, Hotbot, Lycos, Northernlight, and Yahoo for comparative study.

Each researcher submitted a total of five questions in a range of subject areas to each of the eight selected search engines, totaling 2,080 individual searches in 260 search panels of eight search engine trials. Data was collected in the following categories on the first 20 unique citations viewed in the search output lists from the engines:

1) an index of relative recall based on the actual or estimated recall reported by the search engine
2) the number of direct hits among the first 20 unique citations viewed
3) the number of false coordinations among the first 20 unique citations viewed
4) the number of citations to websites with duplicate content
5) the number of citations to websites resulting in failed views
6) the depth to the first solid hit among the citations in the search output list

The aim of the research was to identify the engines that might best meet the needs of a library patron. While, on the whole, the search engines performed equally well on a number of parameters tested, it was found that engines differed most significantly in:

1) the percent of relevancy in results from direct hits
2) the depth to the first solid hit
3) the number of duplicate citations delivered
4) the number of citations which resulted in failed views

A discussion and summary of the results, conclusions and recommendations for further research are included.

1. BACKGROUND

1.1 Overview

This project builds on previous work conducted by classes in the Department of Library and Information Studies of the School of Education at the University of North Carolina (UNC) at Greensboro under the direction of Dr. Randy D. Ralph. Six of the top eight global World Wide Web (WWW) search engines identified in the previous comparative testing in 1997 as part of an Indexing and Abstracting course (WWW Search Engine Test Methods, available at URL http://www.netstrider.com/search/methods.html) and in 1999 as part of a course in library automation (Computer-Related Technologies in Library Management, by Randy D. Ralph and John W. Felts at URL http://library.uncg.edu/search/ were again selected for comparative benchmark testing, this time using the fall 2000 Computer-Related Technologies in Library Management classes (LIS 645), meeting at UNC's Asheville, Charlotte and Greensboro campuses. Each of the fifty-two students devised five (5) search queries in diverse subject areas and genres in order to gauge the overall performance of the eight selected search engines. In a departure from the earlier study, advanced search queries were presented to the search engines using their own advanced search interfaces, rather than the simple default interfaces. The search engines selected were Altavista, Excite, Go/Infoseek, Google, Hotbot, Lycos, Northernlight and Yahoo.

1.2 Rationale

There is still a need for the type of examination performed here. While more and more librarians (among the rest of us) are using search engines, few real statistical analyses, as opposed to popular informal comparisons, have been conducted. Many earlier studies are so old they are outdated, since search engines evolve so rapidly. New studies are underway, but this study builds on earlier research only three years old, expanding the earlier parameters. Moreover, as the Internet becomes more and more commercialized, the need for an unbiased and statistically valid comparison is greater now than ever before. This research can be periodically repeated, taking into account the evolution of the search engines as well as that of the Internet itself.

1.3 Background of Search Engines

Search engines came into existence only after 1994. A search engine is software that searches web sites and indexes found in the World Wide Web, and returns the matches, such as documents compatible with the search query (Lien and Peng) [1]. The software agents crawl the Web, looking for and storing anything not in their indexes, usually entire pages. New material can be added from previously indexed pages that have changed, links to pages not yet indexed and Web site addresses submitted by third parties. Once the index is assembled, a review process eliminates duplicate information, such as multiple versions of a site (mirrors). Some search engines give special status to Web pages that use metatags containing descriptors such as "name," "content" and "keywords," since the page authors go to the trouble of describing what their page contains (Schwartz) [2].

Currently there are more than thirty different engines. There are three broad categories: primary search engines, meta-indexes, and specialized search engines. A primary search engine covers a significant portion of WWW using a random search scheme. This category includes Alta Vista, Excite, Hotbot, Go/Infoseek, Lycos, and the WWWorm, to name a few. A specialized search engine, such as Yahoo, maintains the directory manually (Lien and Peng) [3]

It is possible to estimate the size of the Web on the basis of an overlap among the larger engines. One estimate sets the size of the indexed Web at 320 million pages, but, given documents that can be hidden behind search forms, and differences built into algorithms, this is probably smaller than the true size of the Web (Lawrence and Giles) [4.]

There are three components in a search engine. The first component is the spider, also called the crawler. The spider visits web pages, reads the contents and then follows the links to visit other pages within the same site. In general, the spider returns to the site on a regular basis (say, every month or two) for updating. The index, sometimes also called the catalog, is the second component of the search engine. It keeps all the information generated by the spider. It contains a copy of every web page the spider has visited. The final component is the search software. This software screens through all the pages contained in the index to find possible matches to a search query. It also ranks the matches in the order of relevance (Lien and Peng) [5].

The basic premise of relevancy searching is that results are sorted, or ranked, according to certain criteria. Criteria can include the following: number of terms matched; proximity of terms, location of terms within the document, frequency of terms (both within the document and within the entire database), document length, and other factors. The exact "formula" for how these criteria are applied is the "'ranking algorithm' which varies" among search engines (Courtois and Berry) [6].

1.4 Literature Review

There are many articles that rate features of search engines, but remarkably few recent papers have evaluated the performance of search engines. Many are expository; they describe the search engines' features, and often as not they only compare a relative handful. The marketing departments of the search engines' parent companies develop many more to appeal to advertisers. Copee discusses in the 2000 study how Web sites of companies can climb the search engine rankings on the Internet by carefully packaging their content. He describes the difficulties of establishing a high search engine ranking and provides guidelines in creating the content of a site (Copee) [7].

Others have attempted to evaluate selected search engines on the basis of the retrieval yielded by various searches. Although most provide good descriptions of the engines under investigation, earlier studies (before 1996) fall short of executing a significant number of searches in order to conclude which engine is the most accurate or efficient (Tomaiuolo and Packer) [8]. Martin Courtois[9] and colleagues devised a creative and valid approach in 1995: Using only three sample questions, the authors identified benchmark Web resources they expected the engines to return in a results list (Tomaiuolo and Packer) [10.] This is continuing, and the above-mentioned University

of North Carolina Greensboro study is only one of a growing number. The trend to put a larger number of questions to a greater sampling of search engines is growing.

Some of the earlier studies were very informative. Gregg Notess[11] (1995) compared Infoseek, Lycos and Webcrawler and concluded that Lycos performed better than Infoseek, which in turn performed better than Webcrawler, based on three criteria; coverage, precision and currency (Lien and Peng) [12]. A later article by Notess (published in 2000) examines search engine relevance, and Notess concludes that while relevance is improving, "librarians should feel secure in their jobs", as there is so much information on the web to sift through (Notess) [13].

Courtois et al. (1995) submitted queries from three sample questions to search engines. They considered CUI, Harvest, Lycos, Open Text, Webcrawler, WWWorm, and Yahoo and found that Open Text was the best 'with its flexible, powerful search interface and quick response'. [14] Webcrawler, however, offered 'the easiest interface for novices' (Lien and Peng).[15] Courtois and Berry examined search engines (1999), using advanced search queries to examine five major engines. Of more interest to this study, Courtier examined the criteria for testing relevancy ranking, the methodology used in the search, and the effects of terms, proximity, and location (as weighed by algorithms). In this study, Excite and Lycos had the best ranking (Courtois and Berry). [16]

Scoville in 1996 surveyed a wide range of search engines and recommended Excite, Infoseek and Lycos, as these engines are easy to use for information interfaces (Lien and Peng). [17] Tomaiuolo and Packer made a study in 1996 to quantify accurate matches produced by five search engines for 200 subjects. They note that a possible means of improving accuracy and retrieving more useful information may be to have home page creators and search engine developers standardize how searches are made, and what the pages will provide to facilitate those searches (Tomaiuolo and Packer). [18] This is an interesting precursor to Copee's article above, and the development of description tags.

Leighton and Srivastava in 1997 corrected Leighton's earlier study from 1995 comparing Alta Vista, Excite, Hotbot, Infoseek and Lycos. They enlisted fifteen sample questions from a university library and submitted them as search queries. Based on the precision criteria, Alta Vista, Infoseek, and Excite were the top three engines although Lycos and Hotbot tended to perform better on short queries (Lien and Peng) [19]. Lawrence and Giles conducted a study in 1998 that examined the size of the Internet, and compared many search engines based on the percentage of the web they searched, and concluded that the coverage of the search engines vary by an order of magnitude, covering from 3% to 34% of the Web (Lawrence and Giles) [20].

Lien and Peng conducted a study in 1999 that adopted the data envelopment analysis (DEA) method to investigate the efficiency of several search engines. A query on a search engine is modeled as a production process. The input and output vectors are defined and measured accordingly. They also provide an extremely detailed literature analysis (Lien and Peng). [21]

In 1997 Kenk made a study remarkably similar to ours in intent. A comparative analysis was carried out to evaluate just how different search engines produce results for one and the same search task. A "battery" of search strings was used with popular WWW search engines just to

find out which produces the largest number of hits. The strings used fuzzy logic, and not the advanced search features. This benchmark technique was based on a "semantic portfolio analysis". As part of this analysis, a portfolio of semantic phrases or terms (the benchmark) was applied and each term was used in a comprehensive search process using the most popular search engines (Kenk). His study also went back to reexamine the data after 12 months. He found Alta Vista, Infoseek and Excite to be the best engines in his study (Kenk)[22].

King presents an interesting alternative to a general engine search in an examination of specialized search engines (2000), in which he presents information on several specialized Web search engines which can be used for simplified information retrieval, as well as reasons for using specialized search engines such as Achoo or the HealthAtoZ search engine (King) [24].

Zetter, McCraken and Garone present an evaluation of more than 20 Web search engines, directories and expert sites to see which ones produced the best results. Published in September 2000, they prefer the options in Google, the Open Directory Project directory, and the AskMe.com expert site based on the features, ease of use, and percentage of relevant links (Zetter, McCraken and Garone) [25]. Written for PC Magazine, this is among the best for librarians new to finding their way on the Web.

2. METHODS

This research was conducted by 52 graduate students of library and information studies at three campuses of the University of North Carolina (UNC) - Asheville, Charlotte and Greensboro - under the supervision of Dr. Randy D. Ralph of the Department of Library and Information Studies at UNC Greensboro. The research constituted a class project for the fall 2000 course offering LIS 645, Computer-Related Technologies in Library Management. The participants represent a cross-section of the library community in North Carolina and include both practicing professionals and students studying towards library degrees in various specialties, principally school media.

2.1 Research Design

The participants were charged with the development of a detailed research plan both through in-class discussion over the NC-REN television network and participation in Web-based discussion groups supported on a TopClass server at UNC Greensboro and via a class e-mail discussion forum. The instructor divided the class into four groups at three locations for the purpose of writing and to execute the research plan:

> **I. Background/Literature Review - James B. Gibson, Captain.**
> Participants: Karen Barker, Suzanne Harrington, Deanna Herlong, Heather Holley- Hall, Cassandra Hunsucker, Heather Koonts, Clark Nall, Lisa Persinger, and Elizabeth White.
>
> **II. Methodology/Assumptions/Definitions - Benjamin R. Morgan, Captain.**
> Participants: Jennifer Bingham, Annette Brown, Summer Carr, Nancy Daniel, Pat Dunford, Carla Hollar, Ken Miller, Deborah Schillo, Melanie Stallings, and Leisa Stamey.

III. Results/Discussion - Robert J. Mayer and Cynthia Organ, Co-Captains.
Participants: Sarah (Cathy) Cassidy, Camilla Goulet, Robin Holleman, Brenda Kendrick, Beth Lanzy, Don Lineberger, Cornelia Pleasants, Sandy Prete, Eva Putnam, Betsy Sandberg, Carolyn Thomas, Anne Wilhelm and Holly Williams.

IV.Abstract/Summary/Conclusions - Thomas P. Cole, Captain.
Participants: Shirley Baucom, Kathy D'Aurelio, Karri Freeney, Tim Hunter, Norma Jones, Lisa Newburger, Jennifer Prince, Monika Rhue, Anita Robinson, Edith Smith, Michelle Spink, Donna Surles, Rita Vogel and Michael Winecoff.

The participants elected their own research captain(s) for each group with the responsibility for coordinating and facilitating the research and writing effort among the members. Corresponding threaded discussion areas were set up by the instructor within TopClass for use by the participants. Additionally, a triage area styled the "First Aid Center" was established for resolution of questions and issues arising during the research. A "Captain's Forum" was also established to facilitate communication and coordination of effort among the group captains.

The participants designed the research to benchmark the performance of the advanced search interfaces of the most popular independent general World Wide Web (WWW) search engines. The participants identified the engines to be included in the study by polling themselves and by viewing objective data on referrals to a popular WWW domain (www.iconbazaar.com) over a six day period in November, 2000 (see the figure and table below). Eight search engines with advanced search interfaces were selected from among those reviewed: Altavista, Excite, Go/Infoseek, Google, Hotbot, Lycos, Northernlight and Yahoo. Webcrawler, a popular general engine included in previous benchmarking studies, was rejected because it does not offer an advanced search interface.

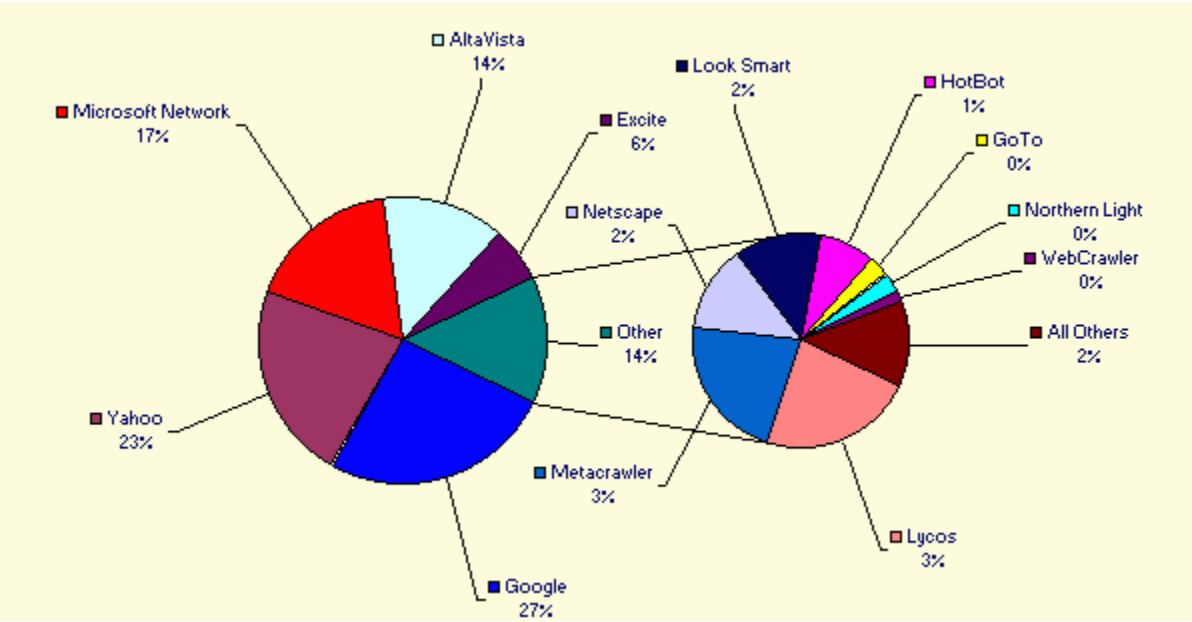

FIGURE 1. Referrals to Domain www.iconbazaar.com from Major WWW Search Engines

11/07/2000 00:00:00 - 11/12/2000 23:59:59		
Wednesday November 15, 2000 - 18:17:40		
Top Search Engines		
Engines	**Searches**	**% of Total**
Google	3,212	25.9%
Yahoo	2,788	22.4%
Microsoft Network	2,159	17.4%
AltaVista	1,714	13.8%
Excite	756	6.1%
Lycos	409	3.3%
Metacrawler	396	3.2%
Netscape	235	1.9%
Look Smart	231	1.9%
HotBot	150	1.2%
GoTo	57	0.5%
Northern Light	56	0.5%
Webcrawler	26	0.2%
All Others	236	1.9%
Totals	**12,425**	**100.0%**

TABLE 1. Search Engine Referral Report for Domain www.iconbazaar.com (6 day scale)

Each participant in the research study was asked to develop five queries in a variety of subject areas for submission to all eight search engines. To ensure that the researchers' questions covered a wide range of subject areas, they were evaluated before the research began. If a set of questions did not cover a sufficient range of topics, individuals were encouraged to create new, more varied questions. The figure below shows the distribution of topics covered in the questions adopted:

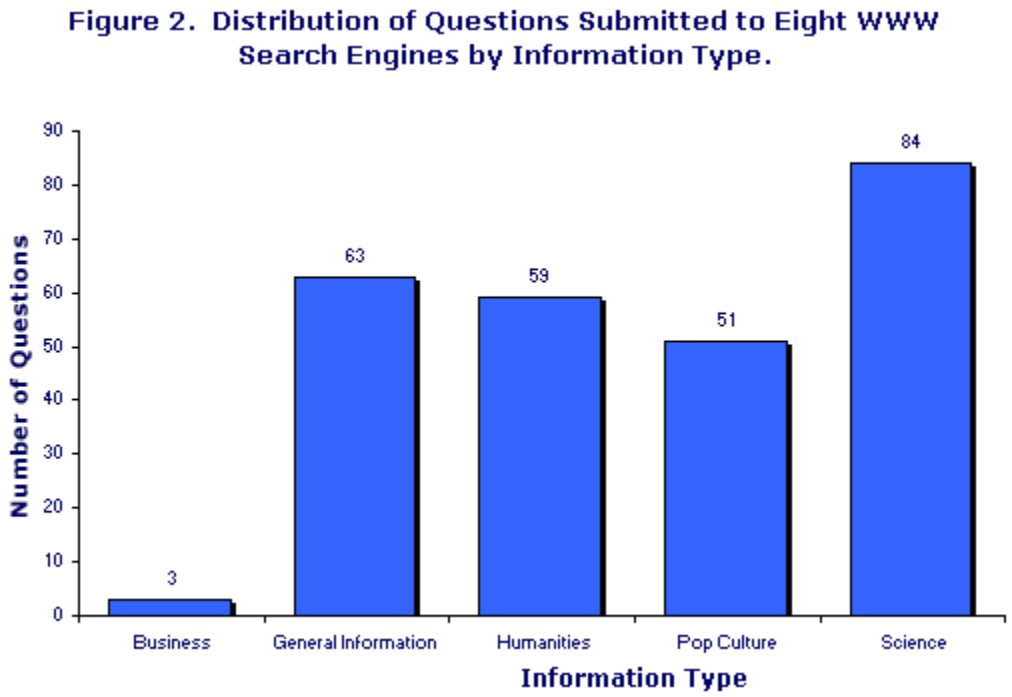

FIGURE 2. Distribution of Questions Submitted to Eight WWW Search Engines by Information Type.

The fifty-two (52) participants submitted a total of 260 panels of 8 searches to the search engines for a grand total of 2,080 individual searches within the period of the study - approximately five weeks. To make the research process more uniform and reduce the chances that search engines queried later might have an advantage over those queried earlier, with the understanding that both researchers and search engines "learn" over time: 1) all results for a single question were recorded within 48 hours; and 2) the research questions were submitted to the engines in random order.

The array of questions submitted represents a sizeable base for statistical analysis of the results. Participants submitted all queries to the advanced interfaces of the selected search engines using retrieval strategies as sophisticated as their searching experience and the capabilities of the interfaces would allow. Any and all search engine interface features that would maximize the effectiveness of the search were employed. There was no systematic attempt to train searchers how to use the engines effectively as it was thought this might introduce a bias into the results which might invalidate their extrapolation to the general population.

Search results were limited to retrieval of citations from the WWW for those engines which included special collections, categories or other output formats in order to fairly test relative database content and quality.

Participants collected data in seven (7) categories on the first 20 unique citations delivered by the search engines in response to the queries submitted as follows:

1. Recall Index - an ordinal scale based on the actual or estimated recall reported by the engines:

7 - unreported
6 - > 1,000,000
5 - 100,001 - 1,000,000
4 - 10,001 - 100,000
3 - 1,001 - 10,000
2 - 101 - 1,000
1 – 0 - 100

2. Citations Viewed - the total number of citations necessary to be viewed in order to evaluate 20 unique citations.

3. Direct Hits - the number of citations to websites delivered in the search output which were directly and visibly relevant to the query posed.

4. False Coordinations - the number of citations to websites delivered in the search output which were not directly relevant to the query posed but which were correctly retrieved by the engines on the basis of the query posed.

5. Duplicates - the number of citations to "mirror" websites with duplicative content delivered in the search output.

6. Failed Views - the number of citations delivered in the output to websites which could not be viewed for any reason, including but not limited to, browser timeouts, 404 not found, server errors, network errors, etc.

7. Depth to the First Hit - the depth in the output list of citations delivered by the engines to the first direct hit, designated as "response."

The rationale for restricting the analysis of output from the engines to the first 20 unique citations rests on the notions that the average Internet user is unlikely to follow results through the end of a very long list of output citations and that the search engines employ algorithms to rank the output placing citations with a greater probability of interest toward the top of the output list.

The data were collected and tabulated on a general Excel spreadsheet template developed and tested by the class for use by all participants. Participants communicated interim and final results to the instructor and to fellow participants using threaded discussion groups supported by the TopClass server at UNC Greensboro. A test run using a single common query was performed at the beginning of the research to test both the data gathering template spreadsheet and to prepare the participants for the actual research. This led to a general refinement of methods and fostered a greater understanding of the mechanics of the information collecting and reporting process among the participants. An example of the final data collection spreadsheet instrument is reproduced below for review:

Make NO Modifications to this Template Except for Data	LIS 645 WWW Search Engine Benchmarking Research Results Reporting Spreadsheet Template						
Researcher:							
Question 1:							
Date Completed:				Random Engine: Excite			
Search Engines	Recall Index	Citations Viewed	Direct Hits	False Coordination	Duplicates Found	Failed Views	Depth to First Hit
AltaVista							
Excite!							
Go/Infoseek							
Google							
Hotbot							
Lycos							
Northernlight							
Yahoo!							

FIGURE 3. Portion of the Data Collection Template Spreadsheet.

Note that the template allowed the researchers to record the actual query posed and the date on which the query in all engines was completed. Columns are provided for the entry of the data to be recorded. The template also has a "random engine" cell which identifies the next engine to be searched in random order, which can be refreshed by the researcher.

2.2 Statistical Analyses

Template spreadsheets were gathered electronically from all 52 research participants. They were incorporated by direct reference to each of the 52 external files into a unified master summary data spreadsheet file in Microsoft Excel. The stereotypic layout of the 52 individual template spreadsheets permitted this approach. All data analyses were performed on the master summary spreadsheet.

Data collected in each category were organized into separate worksheets within the master summary spreadsheet and subjected to 1-way analysis of variance (ANOVA) using the Dunnett test at the 99th percentile as performed by the "Analyse-It" data analysis plug-in for Microsoft Excel.

Data collected in each category were compared by responses from the individual search engines against the average response computed for all search engines in order to reveal significant differences, if any, among the performance of the search engines in each category. Graphics were generated within Excel for each category of data collected and provide the basis for the

figures presented in the results section. Significant variations in performance, as revealed by the Dunnet test, were included in each graphic representation along with error bars based on 2 standard deviations around the calculated means for each data category.

Data were also analysed for each category by individual participant against the average response for all participants in order to reveal any significant departures of researchers' data collection and reporting performance from the general group performance in each category of data collected. Significant departures from the general group performance were noted for discussion. Descriptive statistics were adduced for the general performance of the researchers as a group, as well, in order to reveal closeness of fit to the expected normal distribution among researcher performance.

3. RESULTS & DISCUSSION

Table 1 (below) summarizes the cumulated data for 2,080 searches submitted to the 8 selected WWW search engines in 260 search panels by the 52 research participants. Data presented within each category represent the averages of 260 search panels with The exception of category "Failed Searches."

Search Engines	Recall Index	Citations Viewed	Depth to the First Hit	Direct Hits	False Coordinations	Duplicates	Failed Views	Failed Searches
AltaVista	2.5	21.5	3.5	25.5%	60.9%	4.8	3.0	2.5
Excite!	2.7	23.3	2.3	36.1%	53.7%	17.6	2.4	2.0
Go/Infoseek	2.1	18.4	2.2	31.0%	52.4%	6.2	1.5	2.6
Google	2.9	21.3	1.9	48.9%	45.2%	7.1	1.0	1.1
Hotbot	2.9	20.8	2.3	37.2%	53.5%	5.8	0.8	1.8
Lycos	2.8	22.9	2.3	36.8%	53.0%	11.5	2.5	2.0
Northernlight	3.0	21.5	2.1	40.6%	50.8%	5.8	1.3	1.7
Yahoo!	2.4	19.4	1.9	44.1%	50.2%	2.7	0.9	1.1

Table 2. Summary of Results

The raw data from which these summary data were adduced were subjected to Analysis of Variance (ANOVA) using the Dunnett Single-Tailed Test at the 99% confidence level in order to determine significant deviation of the observed values from the mean responses for the total population. The Standard Error around the mean of each average was computed and used to plot the error bars at 2 times the S.E.

3.1 Citations Viewed

The figures presented in Figure 4, below, show the average number of citations which were required to be viewed in order to evaluate 20 unique citations among the responses to queries submitted to the search engines studies. The research design called for participants to view 20 unique citations in order to more accurately gauge search engine performance and database quality.

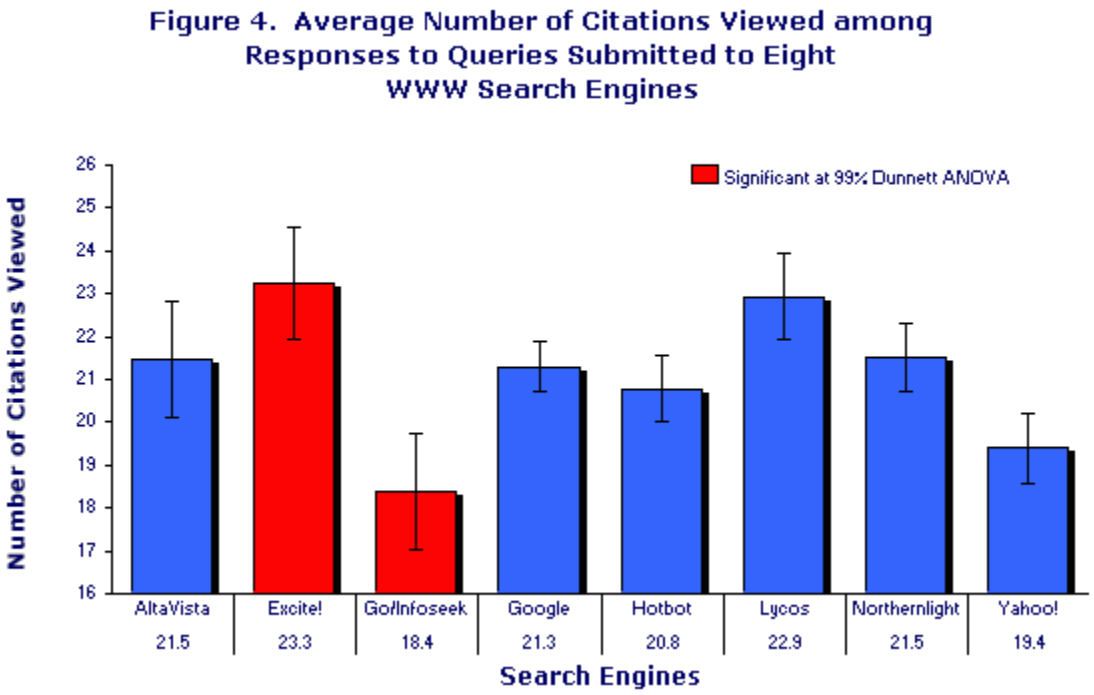

FIGURE 4. Average Number of Citations Viewed Among Responses to Queries Submitted to Eight WWW Search Engines

These figures indicate that, on average, the majority of search engines required viewing slightly more than 20 citations in order to evaluate 20 unique citations. This is in line with the observation, reported below, that the engines deliver, on average, roughly one citation to websites with duplicate content per search. This measure is a rough indicator of both relative database quality and accuracy in reporting of recall figures by the search engines.

Analysis of variance using the Dunnett Test at the 99th percentile revealed that the Go/Infoseek engine delivered, on average, significantly fewer than 20 citations to websites to be viewed. This is consistent with frequent informal observations expressed by participants during the course of the research that the Go/Infoseek engine reported recall figures greater than the number of citations which were actually presented by the engine in the search output. The analysis also revealed that the Excite search engine required viewing significantly more citations in order to evaluate 20 unique websites.

3.2 Recall Index

The data presented in Figure 5, below, are a measure of the number of citations any given search could adduce from a search engine's database. Recall was recorded on a nearly logarithmic ordinal scale from 1 to 7 (see Methods).

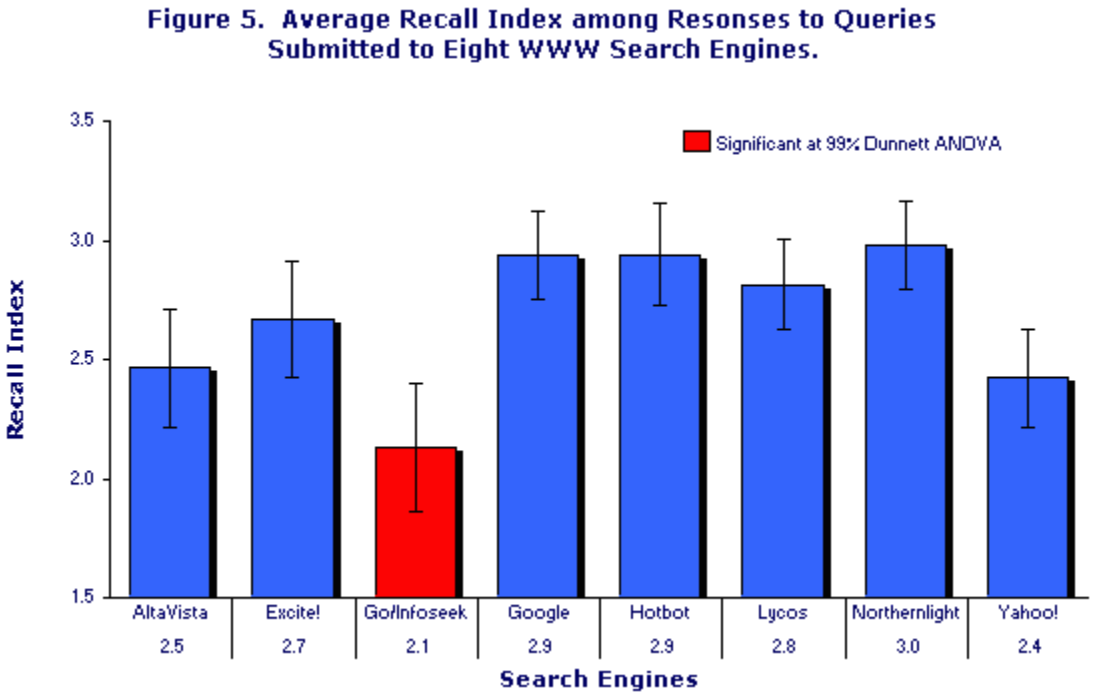

FIGURE 5. Average Recall Index Among Responses to Queries Submitted to Eight WWW Search Engines

The majority of the engines produced results that cluster in the range from 2.47 to 2.98. This corresponds to a recall of from approximately 1,500 to 10,000 citations per search. Analysis of variance using the Dunnett Test at the 99th percentile reveals that the Go/Infoseek engine retrieved, on average, significantly fewer citations than the other engines with an average recall index of 2.13, or, approximately 1,200 citations per search.

3.3 Response

Response was recorded as the "depth to the first hit" in the list of citations output for each search result. Figure 6, below, shows the average response computed for each search engine from all queries submitted.

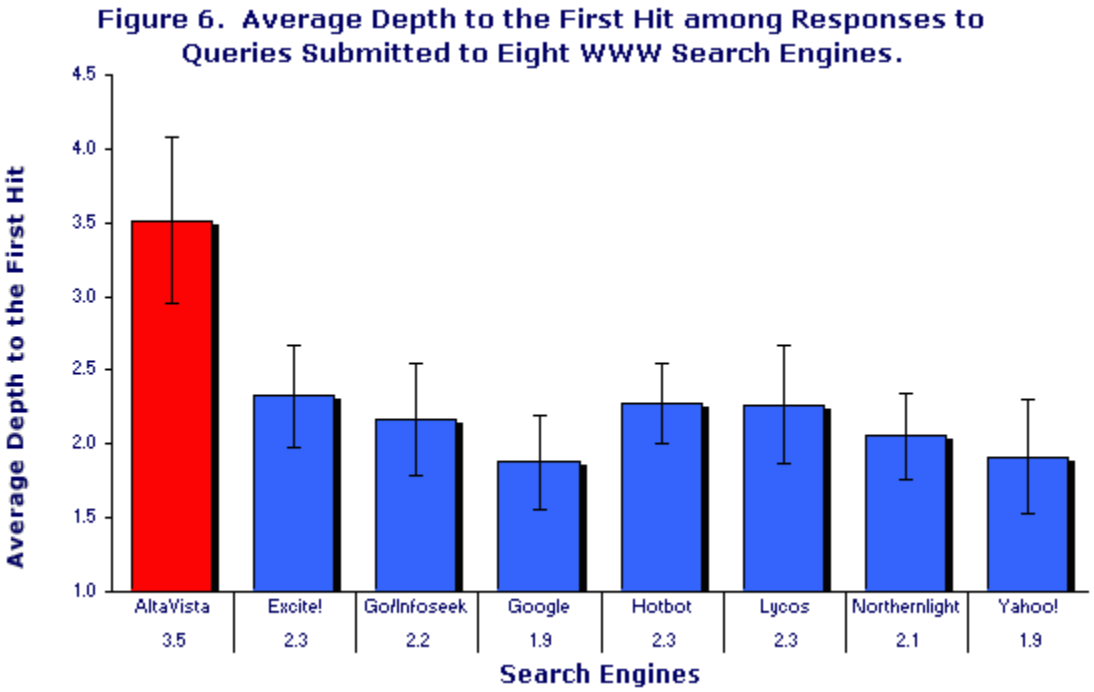

FIGURE 6. Average Depth to the First Hit Among Responses to Queries Submitted to Eight WWW Search Engines

These data clearly show that, for the most part, the average response computed for the search engines lie at an average depth from 1.88 to 2.32 citations. No significant differences were noted among the majority of the search engines studied with the single notable exception of the AltaVista search engine for which the average depth to the first hit was significantly higher. On average the first direct hit found in results from all search engines except AltaVista can be expected to appear at a depth of from 2 to 3 citations in the search output list.

An informal survey of research participants conducted by email subsequent to the compilation of search results revealed that about 81% did not use the "sort by" feature of AltaVista's advanced interface which ranks results according to terms provided by the searcher. Some searchers reported confusion over AltaVista's lack of instructions for using this feature, whereas others questioned the fairness of using a feature not found in the advanced search features of other search engines. This may account, in part, for AltaVista's relatively poor showing.

3.4 Relevancy

Total relevancy is calculated by adding the number of direct hits (websites that fully and visibly answer the search inquiry on the first page) and false coordinations (websites that include all of the required search terms but do not answer the search inquiry), and dividing that result by the number of unique citations viewed and expressing the result as a percentage. The three figures below illustrate how the eight search engines selected for this research performed in terms of relevancy.

Figure 7, below, presents average relevancy from direct hits. Figure 8 presents the average relevancy from false coordinations. Figure 5 presents the average of total relevancy derived as the sum of relevancy from both direct hits and false coordinations.

Figure 7. Average Relevancy from Direct Hits among Responses to Queries Submitted to Eight WWW Search Engines.

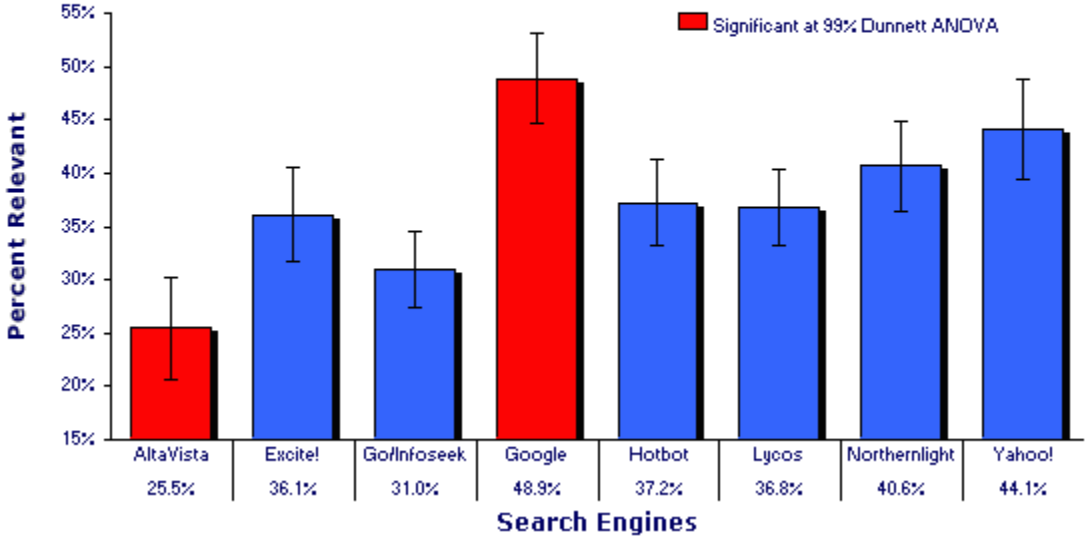

FIGURE 7. Average Relevancy from Direct Hits among Responses to Queries Submitted to Eight WWW Search Engines

Figure 7, above, presents the average percent of relevancy from direct hits produced by each of the eight search engines. As already stated, this study assumes that a "direct hit" is a website that the individual researcher deemed successful in providing a clearly visible answer to the search query. In that context it should be noted that the accuracy of the answer provided was not germane to the determination of relevancy and was not considered. The average relevancy from direct hits from the AltaVista and Google engines differed significantly from that calculated for the remaining search engines. The Google search engine delivered a significantly higher percentage of relevancy from direct hits. The AltaVista search engine delivered a significantly lower percentage of relevancy from direct hits. Differences among the averages computed for the percent relevancy from direct hits among the remaining 6 search engines were not statistically significant and fell in the range from 31.0% to 44.1% overall.

Descriptive statistical analysis of the researchers performance on this measure revealed that the entire group fell well within a normal distribution. Responses from only a few researchers fell significantly above or below the overall average of responses for the group by search engine. This observation gives greater credence to the averages for relevance from direct hits as reported above. The observed differences are more likely a result of variability among the engines than among variability in the way the researchers evaluated results and recorded data.

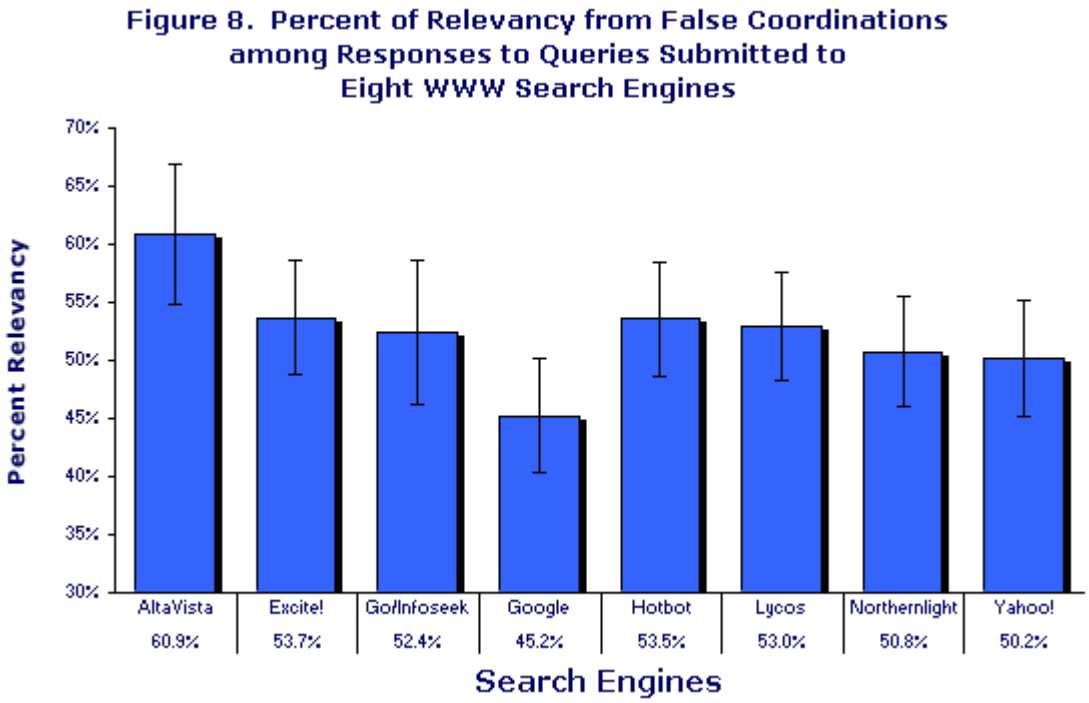

FIGURE 8. Percent of Relevancy from False Coordinations among Responses To Queries Submitted to Eight WWW Search Engines

Figure 8, above, shows the average percent of relevancy from false coordinations produced by each search engine in response to the queries submitted. While some variability is evident from the averages shown above, analysis of variance (ANOVA) indicates no statistically significant variation among these eight search engines with regard to percent relevancy from false coordinations.

Underlying variability in the responses recorded by the research participants may have obscured statistically significant differences. Descriptive statistical analysis of the researchers responses showed that closeness of fit to a normal distribution was poor. The distribution was bimodal with one subgroup reporting consistently low values and another reporting consistently high values. This is the putative source of variability in the data reported here. The researchers were not a unified group in their approach to evaluating the search engines in this measure.

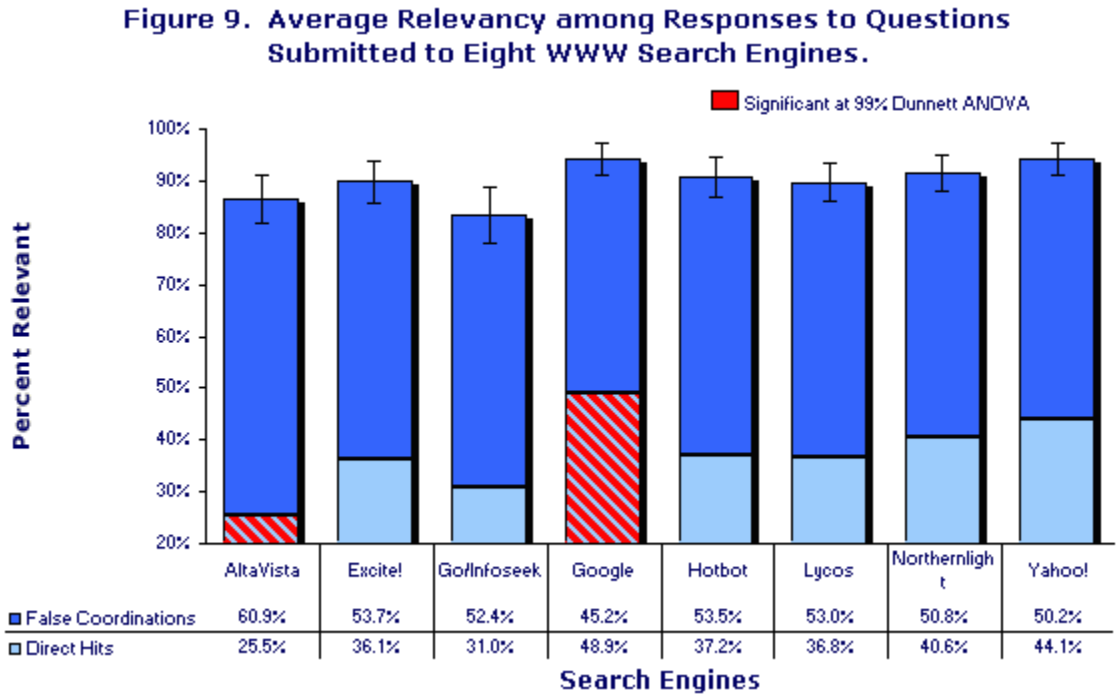

FIGURE 9. Average Relevancy among Responses to Questions Submitted to Eight WWW Search Engines

Figure 9, above, shows the total relevancy computed among responses to queries submitted to the search engines studied. Distinguishing a direct hit from a false coordination calls for a subjective evaluation on the part of the researcher. This inherent subjectivity may have led to the variability in the observed averages for these components, as described above. Determination of total relevancy, however, is a more objective process and provides a real measure of overall relevancy among responses by the search engines. Descriptive statistical analysis of researcher performance on this measure revealed that the group fell within the normal distribution with very few participants reporting values significantly above or below the averages computed for the group. Analysis of variance reveals no statistically significant differences among total relevancy in responses to queries submitted to the eight search engines. Observed percentages of relevancy from all sources were above 85% for all search engines. These observed percentages of relevancy are in line with recent research findings indicating an overall "success rate" of 81% for search engines. (Sullivan, 2000, "NPD Search and Portal Site Study," SearchEngineWatch.com, Available WWW: http://searchenginewatch.com/reports/npd.html.).

3.5 Duplicates

For the purpose of this research, duplicates were defined as citations to "mirror" websites that display essentially identical content in closely similar layouts exclusive of placement of ad banners and without regard to differences in the target URL. The number of duplicates delivered in the result of a query submitted to a search engine is also a relative measure of database quality.

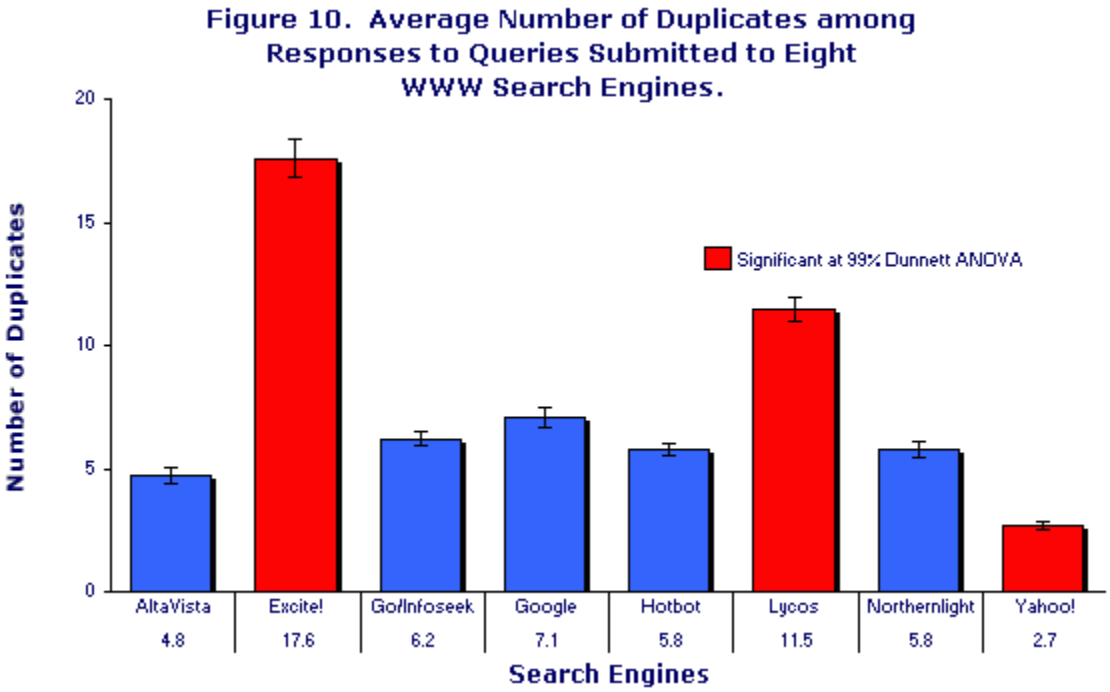

FIGURE 10. Average Number of Duplicates among Responses to Queries Submitted to Eight WWW Search Engines

It is clear from the data shown in Figure 10, above, that the majority of the search engines tested display on average approximately one duplicate per search. Analysis of variance reveals that Yahoo stands out as delivering on average the lowest number of duplicates per search. The analysis also reveals that two search engines, Excite and Lycos, deliver on average more duplicates per search than the other engines.

Descriptive statistical analysis of the performance of the group revealed that the performance of all participants fell well within the normal distribution and that data reported by no participant in the research fell significantly outside the average computed for the group by search engine. These results can therefore be viewed as relatively objective and real measures of database quality among the search engines studied.

3.6 Failed views

Failed views were defined as citations delivered in the output from the search engines linked to websites which could not be viewed for any reason, including but not limited to, browser timeouts, 404 not found, server errors, network errors, etc. Failed views provide a direct relative index of database quality among the search engines evaluated.

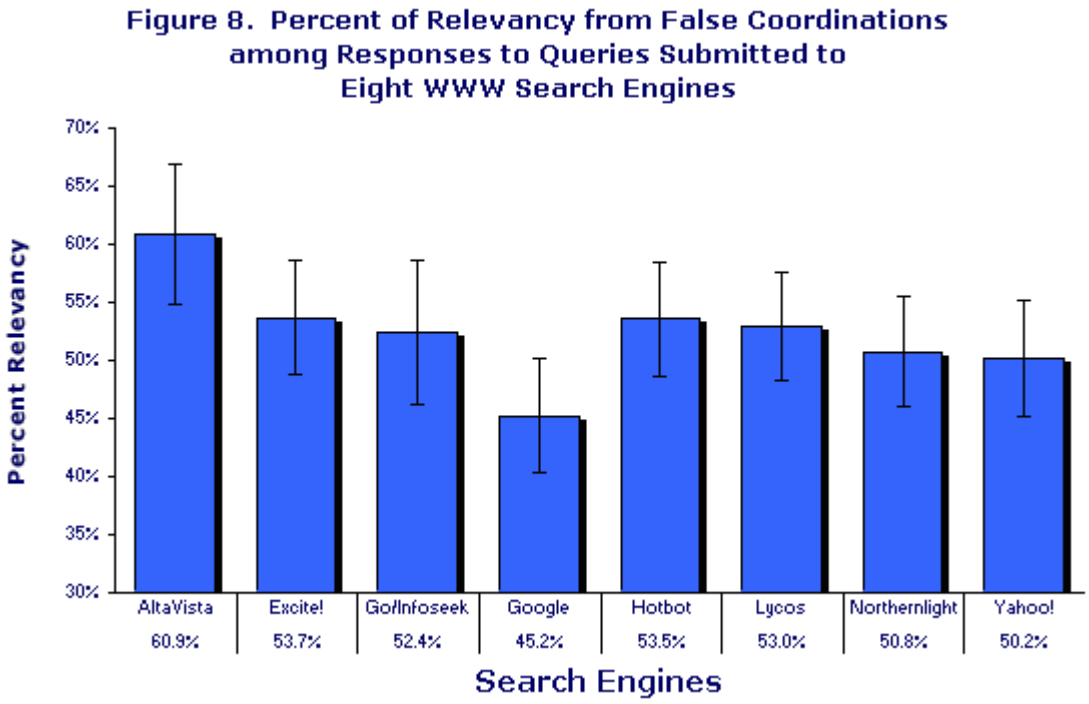

FIGURE 11. Percent of Relevancy from False Coordinations among Responses to Queries Submitted to Eight WWW Search Engines

Figure 11, above, shows the average number of citations leading to failed views in the search output among responses to queries submitted to the search engines studied. Analysis of variance on the average number of failed views per search revealed several significant statistical differences among the selected search engines. Hotbot and Yahoo were shown to deliver significantly fewer citations leading to failed views than the other engines. AltaVista and Lycos, on the other hand, were shown to deliver significantly more citations leading to failed views.

Descriptive statistical analysis of the performance of the group on this measure revealed that the researchers fell well within a normal distribution and that results recorded by no researcher fell significantly outside the range of responses of the group. These data can, therefore, be viewed as relatively objective and real reflections of search engine database quality.

3.7 Failed Searches

Failed searches are those which elicit no response from the search engine at all. They may provide an indication of relative database inclusivity and depth.

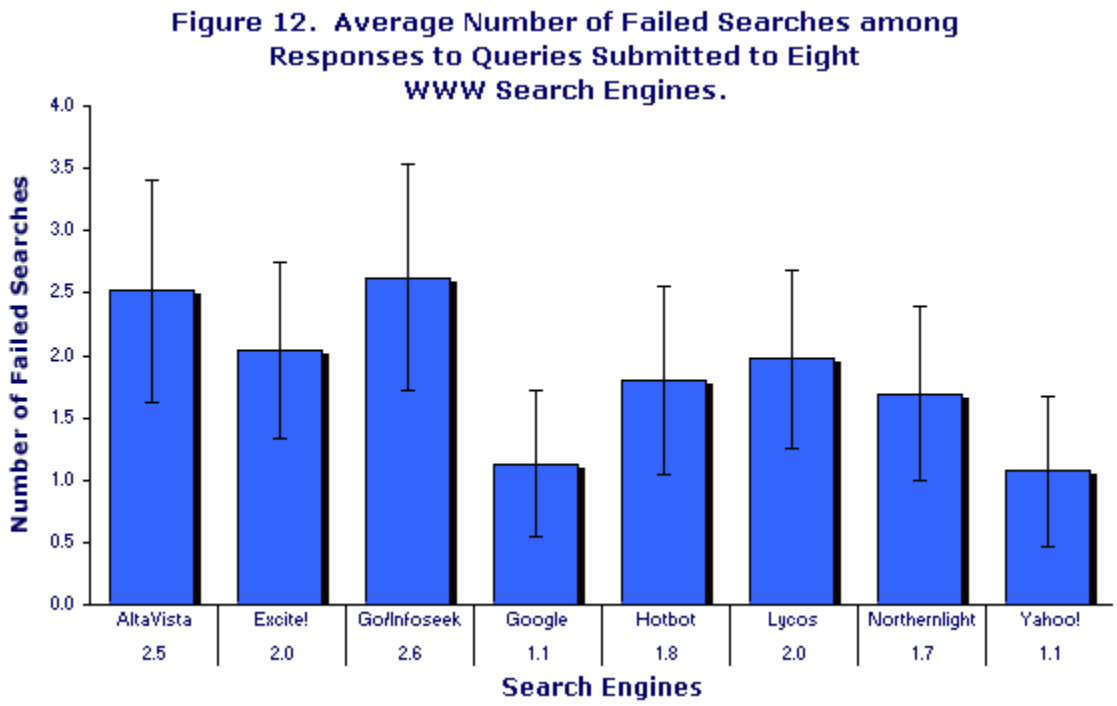

FIGURE 12. Average Number of Failed Searches among Responses To Queries Submitted to Eight WWW Search Engines

Figure 12, above, shows the average number of searches to which the search engines failed to respond at all. Considering that 260 searches were performed, in toto, by all participants using each search engine, the failure rates observed are remarkably low, generally less than 1%. Because the figures are so low relative to the total number of searches performed analysis of variance failed to reveal any significant differences among the average failure rate among the search engines tested. What can be said is that the average number of failed searches varies from a low of 1.1 (Google) to a high of 2.6 (Go/Infoseek) and that there does not appear to be significant variation among the responses.

4. CONCLUSION

This study focused on six (6) measures of performance:

1. Recall (recorded as an ordinal index of recall)
2. Relevancy
    a. from direct hits
    b. from false coordinations
    c. total relevancy from all sources
3. Response (depth in the search output to the first hit)
4. Duplicates (citations to websites with mirrored content)
5. Failed Views (citations to websites which could not be displayed)
6. Failed Searches (queries which elicited no response)

among eight (8) primary search engines on the WWW:

1. AltaVista
2. Excite!
3. Go/Infoseek
4. Google
5. Hotbot
6. Lycos
7. Northernlight
8. Yahoo!

The data presented here indicate that there are no significant differences among performance of the search engines selected for the study with regard to 1) recall; 2b) relevancy from false coordinations; 2c) relevancy from all sources; and, 6) number of failed searches. Significant differences in performance were encountered only with regard to 2a) relevancy from direct hits; 3) response (depth to the first hit); 4) duplicates, i.e., citations to websites with mirrored content; and, 5) failed views.

The data demonstrate that the search engines selected for this study represent a fairly homogeneous group with regard to their overall search performance as elucidated by recall and relevancy. The Google search engine seems to have significantly outperformed its peers with regard to providing more generally relevant search results from direct hits while the AltaVista engine significantly underperformed relative to its peers on the same measure. This would indicate that the Google search engine enjoys a slight advantage over its peers with regard to retrieval of relevant information.

The data show that with regard to measures of database quality there are significant differences among the performance of the search engines. The Excite and Lycos search engines delivered significantly more citations to websites with mirrored content than their peers. The Yahoo search engine delivered the fewest citations to duplicate websites.

The AltaVista and Lycos search engines delivered significantly more citations to websites which could not be displayed while the Hotbot and Yahoo search engines delivered significantly fewer citations to failed views than did their peers. The data taken together indicate that, in general, the overall quality of the Lycos database is generally lower than that of its peers. On at least one measure of database quality it appears that the databases of the AltaVista and Excite engines are lower than that of their peers. The Yahoo search engine excels on two of the measures of database quality over its peers while the Hotbot search engine excels on one measure.

These conclusions are summarized in the table below where it is indicated whether the average performance of each search engine was better or worse than that of its peers for each of the measures of performance where significant differences among the search engines were observed:

Search Engines	Recall	Relevancy from Direct Hits	Response	Duplicates	Failed Views	Failed Searches	Positive Performance	Negative Performance
AltaVista		Worst	Worst		Worst			3
Excite				Worst				1
Go/Infoseek	Worst							1
Google		Best					1	
Hotbot					Best		1	
Lycos				Worse	Worse			2
Northernlight								
Yahoo				Best	Better		2	

TABLE 3. Summary of Relative Performance

Overall, AltaVista received the largest number of indicators of poor performance relative to its peers, followed closely by the Lycos engine. Go/Infoseek failed significantly with regard to recall. Excite failed with regard to number of citations to duplicate websites. The Yahoo engine excelled in two measures of performance while the Google and Hotbot engines excelled in one measure. Significantly, the Google engine consistently delivered the highest percentage of direct hits per search performed. The Northernlight search engine fell within the general range observed among its peers on all measures of performance. No significant differences were noted in the performance of the search engines with regard to the average number of failed searches observed.

5. APPENDICES

5.1 Research Assumptions

1. Students identified the following assumptions in this research:

2. Every participant is capable of evaluating the output of a WWW search, and recording the data according to developed protocols.

3. This was tested during a trial run in which every participant conducted searches on all eight selected engines using a single query. The trial run revealed simple data collection and procedural errors which were subsequently corrected.

4. This group of researchers, comprised of graduate students in library and information studies represents a cross-section of librarians in training in North Carolina, and perhaps in the United States. Each researcher will bring his or her interests and expertise to the study, for example, by creating the questions to be asked of the search engines. Each will have a slightly different approach to his or her search. This variability reflects the variability among all the users of WWW search engines and the questions they posed.

5. Analysis of the preliminary data gathered during the trial run referred to in research assumption #3 demonstrated that, by and large, the researchers fell into a single homogeneous group. Only the searchers' subjective evaluations of relevancy from direct hits and depth to the first direct hit showed variation from a normal distribution. Data gathered in all objective (empirical) data collection categories showed no significant variation among the researchers.

6. The participants are capable of posing each question to all the search engines being tested within a "reasonable" time frame.

7. The research was conducted within a six-week time frame. All students completed all searches within this window.

8. The participants are capable of evaluating the results of the queries in a consistent and objective manner.

9. Analysis of variance among responses from all participants in the actual data collection phase revealed no significant departures from a normal distribution except with regard to subjective evaluations of relevancy from direct hits and depth to the first direct hit, as expected. Otherwise, the researchers were shown to represent a homogeneous group even though the questions submitted by each were different.

10. The participants understand and are capable of using the features of the advanced search engine interfaces.

11. There was considerable discussion of search engines, in general, with in-class examples of how to conduct effective research, as well as a trial run using a standard query. There was no attempt to skew student performance toward a higher plane of understanding of each individual engine's advanced interface as it was thought this would disallow extrapolation of results to the general population in North Carolina.

12. The participants know enough about the questions being asked to ascertain relevancy among the citations delivered by the engines.

13. Participants were encouraged to ask only questions which they were confident they could evaluate for relevancy among responses.

14. The search results received will reflect a cross-section of the answers available on the WWW, and will not merely reflect the relative search skills of the participants.

15. Searcher skills follow a normal distribution as revealed by analysis of data from the trial run. Since the data being collected is relative, not absolute, this factor will not skew results. The engines were queried only against web content, not against any special collections, faceted categories or other value-added features their databases might offer.

16. Questions used to test the search engines will represent a balanced cross-section of the questions that might be asked of a search engine on any given day.

17. Participants were asked to develop queries in a variety of topics: business, humanities, technology, science and pop culture. Queries were examined for accuracy (principally spelling), overlap and duplication. Figure 2 (above) shows that there was good balance in subject matter presented to the engines among the queries submitted.

5.2 Research Protocols

The following protocols were developed to govern data collection and reporting:

1. A random number generator within the data collection template spreadsheet is used to select the next search engine.

2. Only the advanced interfaces of the appropriate search engines will be queried.

3. Only WWW collections will be identified as the object of the search.

4. The recall index that corresponds to the total number of citations returned by the search engine is recorded. If no citations are returned, a recall index of 1 is recorded. If the search engine does not provide information on the number of citations retrieved then a recall index of 7 is recorded.

5. As many citations as necessary are viewed in order to evaluate 20 unique websites and record appropriate data in the template spreadsheet.

6. For every citation hyperlink that is followed each researcher will enter the following values into the “Data Collection Template Spreadsheet” (see Figure 3.):
   - Add 1 to the "Citations Viewed" column.
   - Ask, does the Web page fail to load? If so, it is recorded as a "failed view" by adding 1 to the "Failed Views" column.
   - Ask, does the page mirror the content of a page already viewed? If yes, then this page and the original make a set of duplicates. If the page is one of a new set of duplicates 2 is added to the "Duplicates Found" column. If the page is one of an already identified set of duplicates then 1 is added to "Duplicates Found" column.
   - Ask, does clearly visible information in this page answer the question? If yes, then 1 is added to the "Direct Hits" column.
   - Search the page source code within the browser for the search terms used if the page is determined not to be a "direct hit." If all terms are located within the source code, particulary in the meta tags or image "alt" tag information then the page is counted as a false coordination. 1 is added to the "False Coordination" column.
   - Record the actual depth to the first hit in the output list in the "Depth to First Hit" column if at least one is found among the first 20 unique citations viewed. If no

direct hit is found, either because the search failed to retrieve any citations or because no direct hits were observed in the output list, a null value (not a zero) is left in the "Depth to First Hit" column.

7. The browser's history is cleared on completion of each search before proceeding to the next query.

REFERENCES

1. Balas, Janet. 1998. "The Importance of Mastering Search Engines." *Computers in Libraries* 18(5): 42.

2. Copee, Todd. 2000. "How to Climb the Search Engine Rankings." *InfoWorld* 22(24): 61.

3. Courtois, Martin P. and Michael W. Berry. 1999. "Results ranking in Web Search Engines." *Online* 23(3): 39.

4. Duval, Beverly K. and Linda Main. 1996. "Searching on the Net: General Overview." *Library Software Review* (Winter 1996): 242-251.

5. Feldman, Susan. 1998. "Where Do We Put the Web Search Engines?" *Searcher* 6(10): 40.

6. Hock, Randolph E. 1997. "Sizing up Hotbot." *Online* 21(6): 24.

7. Kenk, Gerhard. 1997. Benchmarking WWW Search Services: A Semantic Portfolio: Analysis of Popular Internet WWW Search Services. [Online]. Available: http://home.t-online.de/home/gerhard.kenk/hpgka500.htm, October 21, 2000.

8. King, David. 2000. "Specialized search engines." *Online* 24(3): 67.

9. Koehler, Wallace. 1998. "Staleness Among Web Search Engines." *Searcher* 6(7): 42.

10. Lawrence, Steve and C. Lee Giles. 1998. "Searching the World Wide Web." *Science* 280(5360): 98.

11. Lien, Donald and Yan Peng. 1999. "Measuring the efficiency of search engines: an application of data envelopment analysis." *Applied Economics* 31(12): 1581.

12. Notess, Greg R. 2000. "Search Engine Relevance." *Online* 24(3): 35.

13. Ralph, Randy D. 2000. Referrals to www.iconbazaar.com Domain from WWW Search Engines. [Spreadsheet]. [Online]. Available: http://www.uncg.edu/~rdralph/LIS645/referrals.xls, October 25, 2000.

14. Ralph, Randy D. and LIS 584B Telecommunications and the Internet class, University of North Carolina at Greensboro. 2000. Search Engine Features. [Unpublished]. Used with permission.

15. Ralph, Randy D. 1997. WWW Search Engines: Indexes, Directories and Libraries. [Online]. Available: http://www.netstrider.com/search/, 15 October 2000.

16. Rappoport, Avi. 1999. "Racing the Engines: The Infonortic Search Engines Meeting, 1999." *Searcher* 7(7): 46.

17. Schwartz, Matthew. 2000. "Search Engines." *Computerworld* 34(19): 78.

18. Tomaiuolo, Nicholas G. and Joan G. Packer. 1996. "An Analysis of Internet Search Engines: Assessment of over 200 Search Queries." *Computers in Libraries* 16(6): 58.

19. Wiggins, Richard and Judith A. Matthews. 1998. "Plateaus, peaks, and promises: The Infonortics '98 Search Engines Meeting." Searcher 6(6): 16.

20. Zetter, Kim, Harry McCraken and Liz Garone. 2000. "How to Stop Searching and Start Finding." *PC World* 18(9): 129.

NOTES

[1]Lien, Donald and Yan Peng. 1999. "Measuring the efficiency of search engines: an application of data envelopment analysis." *Applied Economics* 31(12): 1581.

[2] Schwartz, Matthew. 2000. "Search Engines." *Computerworld* 34(19): 78.

[3]Lien and Peng.

[4] Lawrence, Steve and C. Lee Giles. 1998. “Searching the World Wide Web.” *Science* 280(5360): 98.

[5] Lien and Peng.

[6] Courtois, Martin P. and Michael W. Barry. 1999. “Results Ranking in Web Search Engines.” *Online* 23(3): 39.

[7] Copee, Todd. 2000. “Hot to Climb the Search Engine Rankings.” *InfoWorld* 22(24): 61.

[8] Tomaiuolo, Nicholas G. and Joan G. Packer. 1996. “An Analysis of Internet Search Engines: Assessment of over 2000 Search Queries.” *Computers in Libraries* 16(6): 58.

[9] Courtois and Berry.

[10] Tomaiuolo and Packer.

[11] Notess, Greg R. 2000. “Search Engine Relevance.” *Online* 24(3): 35.

[12] Lien and Peng.

[13]Notess.

[14] Courtois and Berry.

[15] Lien and Peng.

[16] Courtois and Berry.

[17] Lien and Peng.

[18] Tomaiuolo and Packer.

[19] Lien and Peng.

[20] Lawrence and Giles.

[21] Lien and Peng.

[22] Kenk, Gerhard. 1997. “Benchmarking WWW Search Services: A Semantic Portfolio of Popular Internet WWW Search Services.” Available: http://home.t-online.de/home/gerhard.kenk/hpgka500.htm October 21, 2000.

[24]King, David. 2000. “Specialized Search Engines.” *Online* 24(3): 67.

[25] Zetter, Kim, Harry McCraken and Liz Garone. 2000. “How to Stop Searching and Start Finding.” *PC World* 18(9): 129.